Chapter 3

Deep sequencing of the innate immune transcriptomic response of zebrafish embryos to Salmonella infection

This chapter is based on:
Ordas A, Hegedus Z, Henkel C, Stockhammer OW, Butler D, Jansen HJ, Racz P, Mink M, Spaink HP, Meijer AH. Deep sequencing of the innate immune transcriptomic response of zebrafish embryos to Salmonella infection, in preparation

Abstract

Salmonella enterica serovar *Typhimurium* (*S. typhimurium*) bacteria cause an inflammatory and lethal infection in zebrafish embryos. To characterize the embryonic innate host response at the transcriptome level, we have extended and validated previous microarray data by Illumina next-generation sequencing analysis. Comparison of tag-based sequencing (DGE or Tag-Seq) with full transcript sequencing (RNA-Seq) showed a strong correlation of sequence read counts per transcript and an overlap of 241 transcripts differentially expressed in response to infection. A slightly lower overlap of 165 transcripts was observed with previous microarray data. Based on the combined sequencing-based and microarray-based transcriptome data we compiled an annotated reference set of infection-responsive genes in zebrafish embryos, encoding transcription factors, signal transduction proteins, cytokines and chemokines, complement factors, proteins involved in apoptosis and proteolysis, proteins with anti-microbial activities, as well as many known or novel proteins not previously linked to the immune response. Furthermore, by comparison of the deep sequencing data of *S. typhimurium* infection in zebrafish embryos with previous deep sequencing data of *Mycobacterium marinum* infection in adult zebrafish we derived a common set of innate host defense genes that are expressed both in the absence and presence of a fully developed adaptive immune system and that provide a valuable reference for future studies of host-pathogen interactions using zebrafish infection models.

Introduction

In the recent years zebrafish has become widely used as a model for *in vivo* studies of host-pathogen interactions. Zebrafish develop both an innate and adaptive immune system with notable similarities to that of mammals (Traver et al., 2003; Trede et al., 2004). Zebrafish embryos can be exploited to study innate immunity separately from adaptive immune functions, since components of the innate immune system are functional already at the first day of embryogenesis contrary to the adaptive immune system that is not active during the first weeks of zebrafish development (Davidson and Zon, 2004; Herbomel et al., 1999; Lam et al., 2004; Willett et al., 1999). Furthermore, the externally developing and transparent zebrafish embryos are highly suited for real-time analysis of host-pathogen interactions, which can be combined with efficient gene knock-down analysis using antisense morpholino oligonucleotides. It has been demonstrated that the components of the main innate immune signaling pathways are strongly conserved between zebrafish and mammals (Meijer et al., 2004; Stein et al., 2007) and several infection models for studying innate immune response mechanisms in zebrafish embryos have now been developed (Meeker and Trede, 2008). *Salmonella* infections, causing salmonellosis and typhoid fever, are studied in several animal models, of which the best studied is the mouse

model of *S. enterica serovar Typhimurium* infection (hereafter referred to as *S. typhimurium*) (Santos et al., 2001). The opportunity of real time analysis led to the development of a *S. typhimurium* infection model in zebrafish embryos (van der Sar et al., 2003). Intravenous infection of 1 day old zebrafish embryos with *S. typhimurium* strain SL1027 resulted in a lethal infection with bacteria showing intracellular replication in macrophage-like cells as well as extracellular replication in micro-colonies at the epithelium of blood vessels. In contrast, lipopolysaccharide (LPS) mutants of *S. typhimurium* (Ra) were non-pathogenic in zebrafish embryos, similar as in mammalian hosts (van der Sar et al., 2003). Components of the *S. typhimurium* cell wall and motility apparatus trigger innate host defense pathways, including Toll-like receptor (TLR) signaling (Salazar-Gonzalez and McSorley, 2005). A morpholino knockdown analysis of the common TLR-adaptor protein, MyD88, showed that zebrafish embryos lacking MyD88 function lost the ability to clear an infection with the attenuated *S. typhimurium* Ra mutant strain, demonstrating that the innate immune response of the zebrafish embryos involved MyD88-dependent signaling (van der Sar et al., 2006). To characterize the zebrafish embryonic host immune response to *S. typhimurium* wild-type and Ra mutant infection a time-course microarray analysis was performed, showing the induction of genes encoding cell surface receptors, signaling intermediates, transcription factors, and inflammatory mediators, with strong similarity to host responses detected in other vertebrate models and human cells (Stockhammer et al., 2009). A conserved role of zebrafish Toll-like receptor 5 (TLR5) homologs in recognition of Salmonella flagellin was demonstrated (Stockhammer et al., 2009). Furthermore, similar as mammals, zebrafish embryos were shown to employ both MyD88-dependent and MyD88-independent signaling pathways during infection (Stockhammer et al., 2009).

As demonstrated by our previous microarray analysis, the *S. typhimurium*-zebrafish model presents a useful case study for the embryonic innate host response to an inflammatory bacterial infection (Stockhammer et al., 2009). Here we have extended this microarray study by a deep sequencing analysis using the previously described tag-based sequencing method known as digital gene expression (DGE) (Hegedus et al., 2009; t Hoen et al., 2008) also named Tag-Seq (Morrissy et al., 2009). We determined the overlap between deep sequencing and microarray data and report a detailed annotation of the *S. typhimurium*-responsive gene set validated by both analysis methods. Furthermore, we compared the tag-based sequencing approach with full transcript sequencing (RNA-Seq), and based on the overlap between the data sets demonstrate the usefulness of both deep sequencing approaches for transcriptome quantitation during infection. We compared the data with our previous deep sequencing analysis of *Mycobacterium marinum* infection in adult zebrafish and annotated the gene set commonly induced in both infection models. These annotated gene sets provide a valuable reference for future studies using zebrafish infection models.

Results

Identification of Salmonella-responsive genes by tag sequencing

Previously we have performed a microarray analysis of the innate immune response of one day old zebrafish embryos to *Salmonella typhimurium* infection (Stockhammer et al., 2009). Over the first 8 hours of the infection we observed induction of an increasing number of specific gene groups including transcription factors, signaling molecules, and inflammatory mediators. Here we took the 8 hours post infection (hpi) time point for a deep sequencing analysis. Using the digital gene expression (DGE) procedure previously described (Hegedus et al., 2009), hereafter referred to as Tag-Seq (Morrissy et al., 2009), we obtained around 10 million sequence-specific tags from control and infected embryos each. A total of 2471 tag entities showed significantly different expression between the libraries from control and infected fish. Of all significant tags 66% (1630 tags) could be mapped to the UniGene database. The majority of these tags (95%) showed up-regulated expression during infection, while only 5% was down-regulated. Mapping of the significantly changed tags identified a total of 959 different UniGene transcripts when the tags mapping to multiple transcripts were excluded (and 2049 including the tags with multiple mapping). A total of 815 of these transcripts were identified only by tags mapping to the sense strand, 121 transcripts by tags mapping only to the antisense strands, and 23 transcripts collected significant tags mapping to both the sense and antisense strands. Gene ontology (GO) analysis of the up-regulated transcripts with tags mapping to the sense strand showed enrichment of the GO-term response to stimulus ($P<0.01$), consistent with earlier microarray results (Stockhammer et al., 2009).

Validation of Salmonella-responsive genes by comparative analysis of tag sequencing and microarray data

Approximately 60% of the UniGene transcripts that were identified by up- or down-regulated tags in the Tag-Seq analysis (sense strand mapping) were present on the custom Agilent microarray platform used in our previous study (Stockhammer et al., 2009). From those UniGene transcripts, 111 were changed in the same direction (108 up-regulated and 3 down-regulated), 1008 were not significantly changed in the microarray study, and 10 showed a change in the opposite direction between deep sequencing and microarray data (Fig. 1A). Conversely, in the microarray analysis 1254 genes in total were significantly changed (929 up-regulated and 325 down-regulated with fold change >1.5), of which 1133 were not significantly changed in the Tag-Seq analysis. The limited overlap between microarray and deep sequencing data can be explained by differences in methodology, experimental design (e.g. number of parallels), and data processing workflows that are based on different statistical methods to select significant hits. From a biological point of view, the statistical method used to accept expression changes in Tag-Seq analysis as significant (Lash et al., 2000) appeared to be more stringent, as it relied on fold changes of at least a fac-

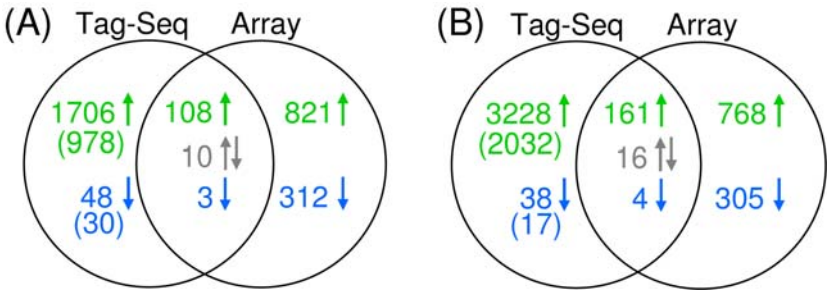

Figure 1. Comparison of Tag-Seq and microarray data of *Salmonella* infection of zebrafish embryos. The microarray data set of 1-day-old zebrafish embryos at 8 hours after infection with *Salmonella typhimurium* is taken from Stockhammer et al. (2009) and compared here with Tag-Seq data of the same samples. (A) Venn diagram showing the overlap of the microarray data set with the Tag-Seq dataset based on statistical evaluation according to Lash et al., (2000). (B) Venn diagram showing the overlap of the microarray data set with the Tag-Seq dataset based on sCTDI calculation. For the transcripts changed only in Tag-Seq analysis (left part of the Venn-diagrams) the numbers in brackets indicate presence on the microarray platform. In the overlapping sectors of the Venn-diagrams the numbers of up-regulated transcripts are indicated in green, numbers of down-regulated transcripts in blue, and numbers of transcripts changed in opposite direction in Tag-Seq and microarray analysis are indicated in grey.

tor 2 for the more abundant tags and over 5 fold change for the lower abundant tags. As an alternative to this statistical method we developed the Cumulative Transcript Detection Index (CTDI) as a means to evaluate differential expression in Tag-Seq datasets. As previously noted (Hegedus et al., 2009), in Tag-Seq analysis the majority of transcripts are represented by more than one tag in the sequence data. The CTDI value reflects the accumulated information from all tags that map to the same transcript, giving increased weight to tags with higher significance in statistical testing (Lash et al., 2000) and giving lower weight to those transcripts where tags are present that change in the opposite direction. We separately calculated the CTDI values for tags mapping to the sense strands (sCTDI) and antisense strands (asCTDI) of the UniGene database transcripts (Fig. 2) and used the sCTDI data for comparison with our microarray data (Fig. 1B). We found that 165 of the transcripts that were differentially expressed according to sCTDI calculation were changed in the same direction in our microarray analysis (161 up-regulated and 4 down-regulated; Supplementary Table 1). Therefore, the responsiveness of these transcripts to *Salmonella* infection in embryos has been confirmed by two independent transcriptome analysis methods.

Validation of Salmonella-responsive genes by comparative analysis of Tag-Seq and RNA-Seq data

Validation of the Tag-Seq data by comparison with microarray data was limited

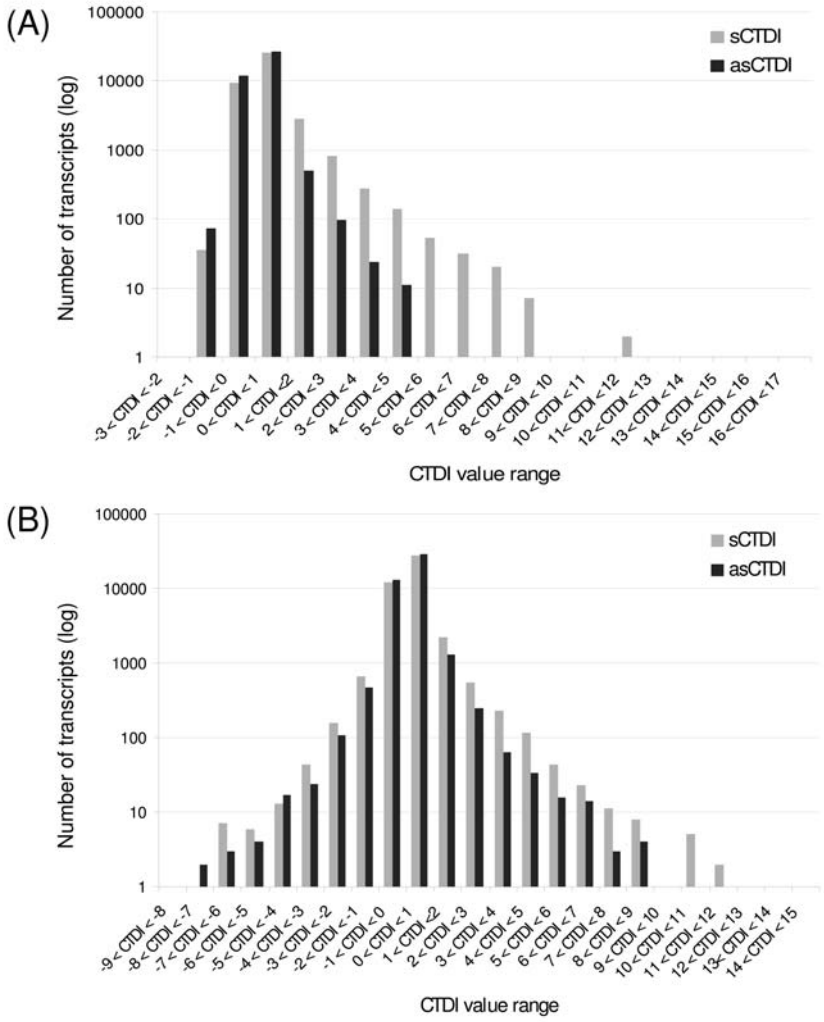

Figure 2. Evaluation of Tag-Seq data by Cumulative Transcript Detection Index (CTDI) calculation. The distribution of the number of transcripts over the range of CTDI values is plotted on a logarithmic scale. Data for tags mapping to the sense (sCTDI) and antisense (asCTDI) strands of the UniGene transcript database are plotted separately. Transcripts represented by at least one tag with a significant expression change based on statistical evaluation according to Lash et al. (2000) have an absolute CTDI value larger than 0.95. For details of CTDI calculation see the Materials and Methods section. (A) CTDI plot of *Salmonella* infection of zebrafish embryos. sCTDI values ranged between -2 and +15 and asCTDI values ranged between -2 and +7 and were in similar range when tags were mapped to the RefSeq (-4<sCDTI<11, -2<asCTDI<4) or Ensembl (-5<sCDTI<10, -3<asCTDI<4) databases. (B) CTDI plot of Tag-Seq data from *Mycobacterium* infection of adult zebrafish (Hegedus et al., 2009).

by the fact that only about one third of all transcripts in the UniGene database were represented on the microarray platform. To extend the validation of our Tag-Seq dataset we therefore performed a comparison with RNA-Seq, in which whole cDNA transcripts are fragmented and deep sequenced. For RNA-Seq analysis, control and *Salmonella*-infected embryo samples from an independent experiment were used. Approximately 15 million reads were obtained for both samples subjected to RNA-Seq. For comparison between RNA-Seq and Tag-Seq the reads were mapped to the Ensembl transcript database based on the Zv8 genome sequence. In both cases approximately half of the sequence reads could be mapped to Ensembl transcripts if no mismatches were allowed. When we allowed 1 mismatch for mapping of the longer RNA-Seq reads (51 nucleotides in RNA-Seq as compared to 17 nucleotides in Tag-Seq) the mapping efficiency could be increased to approximately two-thirds of the total reads. Mapping of RNA-Seq reads identified 85-86% of the known Ensembl transcripts, while 63-69% of transcripts were identified by mapping of Tag-Seq reads (Fig. 3A). Pearson correlations for RNA-Seq versus Tag-Seq libraries of control and infected embryos were 0.84 and 0.81 respectively, which is relatively high especially considering that RNA-Seq and Tag-Seq libraries were constructed from different biological samples (Fig. 3B). As can be observed in the correlation plots, the agreement between RNA-Seq and Tag-Seq data was better for the moderate to highly expressed transcripts than for the low abundant transcripts (Fig. 3B). Statistical evaluation showed that 1244 transcripts were differentially expressed in Tag-Seq analysis and 976 transcripts in RNA-Seq analysis. In both cases more transcripts were up-regulated than down-regulated (Fig. 3C, 3D). However, detection of down-regulated transcripts appeared more efficient with RNA-Seq, where 17% of all differentially expressed transcript were down-regulated, than with Tag-Seq, where the down-regulated transcripts comprised only 3% of all differentially expressed transcripts. Gene ontology (GO) analysis of the up-regulated transcripts of both Tag-Seq and RNA-Seq analysis showed enrichment of the GO-term response to stimulus ($P<0.05$). In total 160 up-regulated and 7 down-regulated transcripts were overlapping between Tag-Seq and RNA-Seq analysis, while 23 transcripts showed changes in opposite directions between the two methods (up in Tag-Seq and down in RNA-Seq) (Fig. 3D). When the Tag-Seq data were evaluated using the above described CTDI algorithm for cumulative detection of tags mapping to the same transcript, the overlap between Tag-Seq and RNA-Seq increased to 233 up-regulated and 8 down-regulated transcripts (Fig. 3E, Supplementary Table 1). In conclusion, Tag-Seq and RNA-Seq were similarly suitable to derive quantitative information on infection-responsive gene expression and the overlap of differentially expressed transcripts between these two deep sequencing methods was slightly better than the overlap between Tag-Seq and microarray analysis (compare Fig.1A,B, Fig.3D,E) .

Annotation of the validated Salmonella-responsive gene set

After collapsing UniGene and Ensembl transcript IDs to single genes, the total set

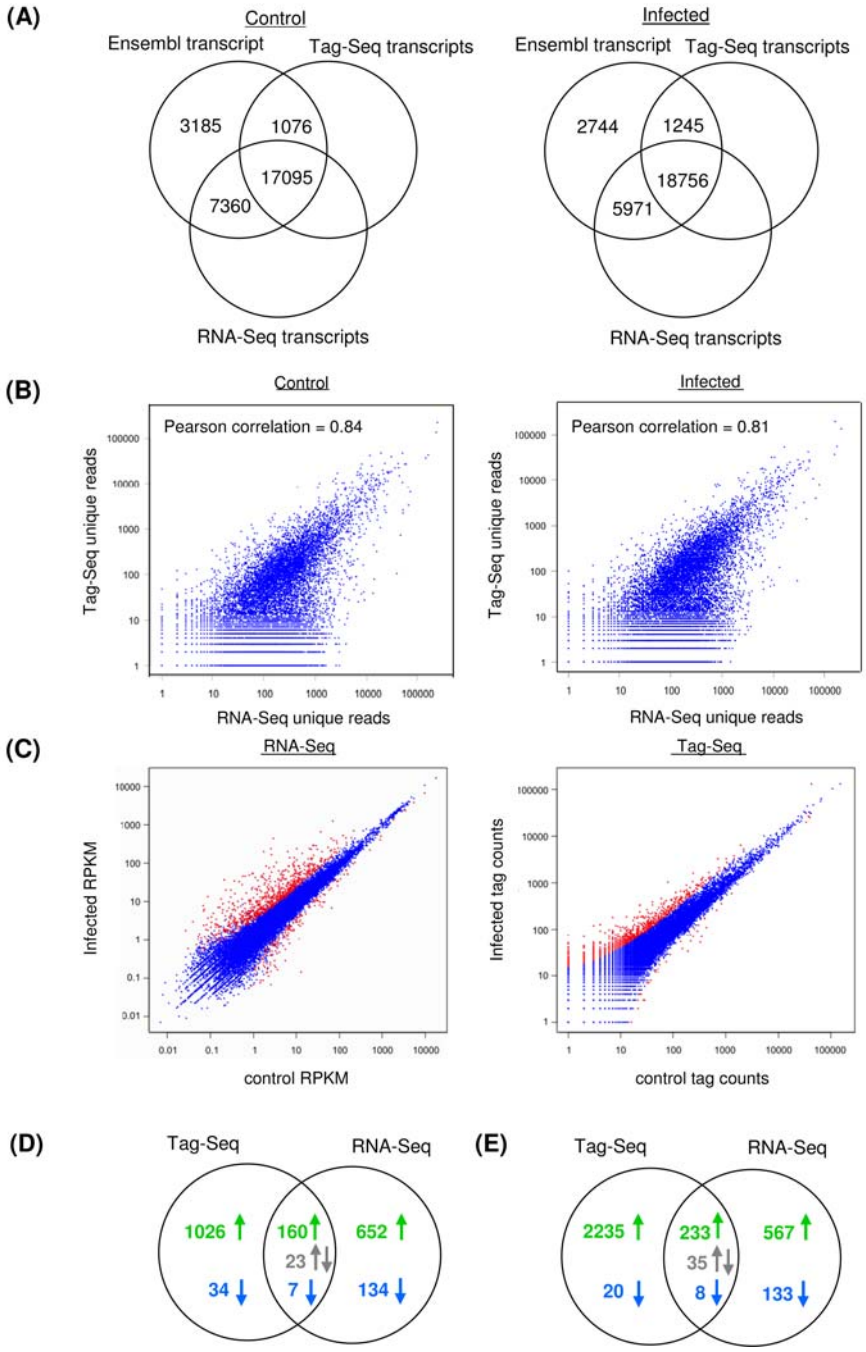
(A)
Control
Ensembl transcript
Tag-Seq transcripts
3185
1076
17095
7360
RNA-Seq transcripts
Infected
Ensembl transcript
Tag-Seq transcripts
2744
1245
18756
5971
RNA-Seq transcripts
(B)
Control
Pearson correlation = 0.84
Tag-Seq unique reads
RNA-Seq unique reads
Infected
Pearson correlation = 0.81
Tag-Seq unique reads
RNA-Seq unique reads
(C)
RNA-Seq
Infected RPKM
control RPKM
Tag-Seq
Infected tag counts
control tag counts
(D)
Tag-Seq
RNA-Seq
1026
160
652
23
34
7
134
(E)
Tag-Seq
RNA-Seq
2235
233
567
35
20
8
133

Figure 3. RNA-Seq analysis of *Salmonella* infection of zebrafish embryos. (A) Efficiency of transcript detection by Tag-Seq and RNA-Seq. The Venn-diagrams show the overlap between all Ensembl transcripts and the transcript identified by mapping of reads from Tag-Seq or RNA-Seq analysis of control and infected embryos. One mismatch was allowed for mapping of the 51 nucleotides long RNA-Seq reads and no mismatches were allowed for mapping of the 17 nucleotides long tag entities in Tag-Seq. (B) Correlation between Tag-Seq and RNA-Seq data. Scatter plots show the counts per Ensembl transcript for both methods. Only reads mapping to a single transcript were included in the analysis. (C) Tag-Seq and RNA-Seq detection of differential expression between control and infected embryos. Scatter plots show the counts per Ensembl transcript (RNA-Seq) or counts per tag (Tag-Seq) for libraries of control versus infected embryos. Transcripts (RNA-Seq) or tags (Tag-Seq) with significantly different expression between control and infected libraries are indicated in red. In RNA-Seq the reads were first mapped to Ensembl transcripts and subsequently differential expression was statistically evaluated according to Lash et al. (2000). The RNA-Seq scatter plot shows RPKM values, which are the total read counts per kilobase per million mapped reads (Mortazavi et al., 2008). In Tag-Seq differential expression was statistically evaluated at the level of the individual tags. (D) Venn diagram showing the overlap of the Tag-Seq and RNA-Seq data sets based on statistical evaluation according to Lash et al. (2000). (E) Venn diagram showing the overlap of the Tag-Seq and RNA-Seq data sets, based on sCTDI calculation for the Tag-Seq data. Tag-Seq reads mapping to the antisense strands of Ensembl transcripts were excluded from all analyses. In RNA-Seq, information on transcript directionality is not obtained and therefore all read mapping data are included in the analyses.

of *Salmonella*-responsive genes confirmed by two transcriptome analysis methods consisted of 317 genes, of which 163 were overlapping between Tag-Seq and microarray analysis, 230 between Tag-Seq and RNA-Seq, and 76 by all three methods of gene expression profiling. This set of 317 genes (309 up-regulated and 8 down-regulated) was taken for a detailed annotation (Supplementary Table 1, Fig. 4). We categorized the differentially expressed genes into 5 categories: 1A - annotated genes previously implicated in the vertebrate immune response (82 up + 4 down), 1B - novel/hypothetical genes with similarity to genes previously implicated in the vertebrate immune response (31 up), 2A - annotated genes not previously linked to the immune response (100 up + 2 down), 2B - novel/hypothetical genes with similarity to genes not previously linked to the immune response (50 up + 2 down), and 3 - genes with unknown function and for which we could not derive any functional prediction (46 up). The genes in the 1A/B and 2A/B categories were ordered by (predicted) functions based on literature and gene ontology data (Supplementary Table 1, Fig. 4). The annotated or novel/hypothetical genes with homology to human immune-related genes (categories 1A/B) together comprised 22 infection-induced genes with functions in transcription activation or repression, including transcription factors of the ATF, AP-1(JUN/FOS), CEBP, ETS, IRF, MYB, MYC, NFκB, and STAT families. Categories 1A/B also contained 21 up-regulated genes involved in immune-related signal transduction pathways, including MAP kinase (MAPK), ERBB2, interleukin

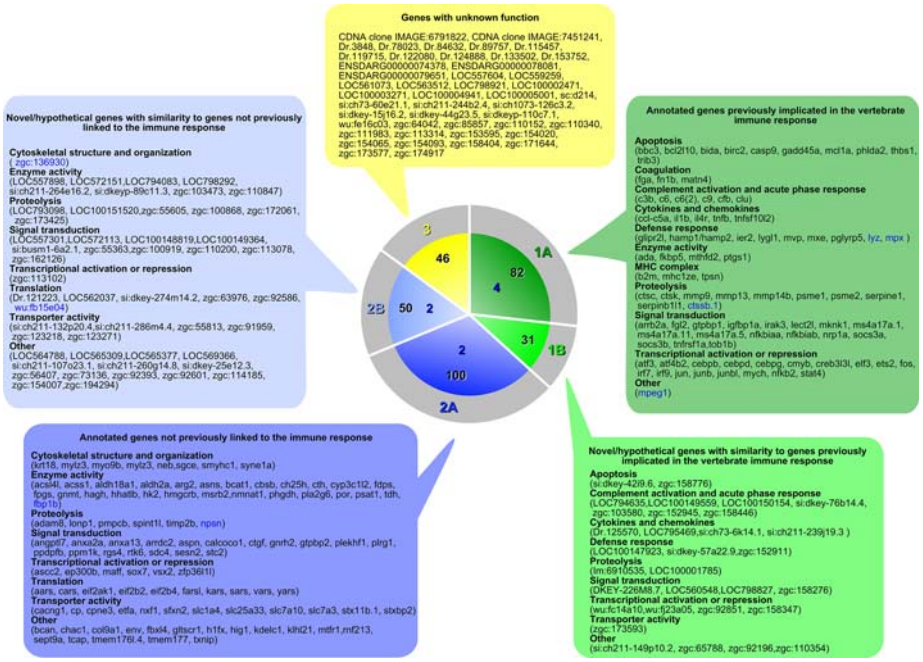

Figure 4. Annotation of the gene set responsive to *Salmonella* infection in zebrafish embryos. Differential expression of genes in the diagram was confirmed by Tag-Seq and microarray analysis, or by Tag-Seq and RNA-Seq analysis, or by all three transcriptome profiling methods as indicated in Supplementary Table 1. Genes were grouped into five categories: category 1A - annotated genes previously implicated in the vertebrate immune response based on GO annotations of the zebrafish genes and their human homologues, on PubMed abstracts and on overlap with the common host response defined by Jenner and Young (2005), 1B - novel/hypothetical genes with similarity to genes previously implicated in the vertebrate immune response, 2A - annotated genes not previously linked to the immune response, 2B - novel/hypothetical genes with similarity to genes not previously linked to the immune response, 3 - genes with unknown function. The number of genes in each category is indicated in the pie diagram and the corresponding genes are ordered by functional groups. Genes up-regulated by *Salmonella* infection (309 genes) are indicated in black, and genes down-regulated (8 genes) are indicated in blue. Gene descriptions and accession numbers are given in Supplementary Table 1.

1 receptor (IL1R) and Toll-like receptor (TLR) signaling. The genes involved in TLR signaling included several negative regulators of the pathway, including *irak3*, *socs3a* and *socs3b*, and the NFκB inhibitor genes *nfkbiaa* and *nfkbiab*, as previously noted in Stockhammer et al. (2009). The signal transduction group also contained 4 members of the membrane-spanning 4-domains subfamily. The up-regulated genes in categories 1A/B further comprised 9 genes with (predicted) cytokine or chemokine activity, 13 genes related to complement activation and the acute phase response, 12 genes

involved in apoptosis and 11 genes with proteolytic functions, including proteasome activator subunit genes, cathepsins, serpins and matrix metalloproteinases (*mmp9*, *mmp13* and *mmp14b*). Furthermore, categories 1A/B contained smaller groups of up-regulated genes encoding MHC complex proteins, coagulation factors, enzymes (e.g. the prostaglandin biosynthetic gene *ptgs1*) and several genes that we classified under defense response, such as the antimicrobial hepcidin (*hamp1/2*) gene, the peptidoglycan recognition (*plyrp5*) gene, the immediate early response gene, *ier2*. The categories 1A/B contained only four down-regulated genes of which two, *mpx* (myeloperoxidase) and *lyz* (lysozyme C), are also clearly linked to the defense response. The down-regulation of these genes might suggest a Salmonella-specific mechanism to counteract host defense. The down-regulated genes further included a cathepsin gene (*ctssb.1*) and the macrophage expressed gene 1 (*mpeg1*), which encodes a perforin-like protein whose precise function is unknown. In contrast, two other cathepsins (*ctsc*, *ctsk*) and an *mpeg1*-like gene (*zgc:110354*) were present in the up-regulated gene set. The 150 up-regulated and 4 down-regulated genes of the 2A/B categories that have not previously been directly linked to immune response were associated to GO terms such as cytoskeletal structure and organization, enzyme activity (including several metabolic genes), translation, and transporter activity (including several solute carriers). The 2A/B categories also contained up-regulated genes associated with proteolysis, signal transduction and transcription factor or cofactor functions, which were not previously linked specifically to host defense.

Comparison of different infection studies

Previously we have used Tag-Seq analysis to investigate the host response of adult zebrafish at the end stage of *Mycobacterium marinum* infection ((Hegedus et al., 2009); chapter 2). The end stage of *Mycobacterium* infection in adult fish is associated with a strong inflammatory response, similar to what we observed during *Salmonella* infection of zebrafish embryos. Therefore, we decided to compare the specific gene groups regulated in both infection studies. To this extent we re-examined our previous Tag-Seq data set from *Mycobacterium* infection by CTDI calculation (Fig. 2B). The CTDI values for tags mapping to the sense strands (sCTDI) of the UniGene database transcripts ranged between –7 and +12, whereas these values ranged between -2 and +15 in the *Salmonella* Tag-Seq data set, indicating that more transcripts were down-regulated in the *Mycobacterium* infection study. This is consistent with our previous analyses that showed a large number of genes encoding metabolic enzymes and muscle proteins to be down-regulated during *Mycobacterium* infection when the fish become strongly emaciated (Hegedus et al., 2009; Meijer et al., 2005). Next, we determined the overlap between the up- and down-regulated transcripts of the *Salmonella* and *Mycobacterium* infection studies (Fig. 5A, Supplementary Table 2). We found 228 and 3 commonly up- or down-regulated transcripts that corresponded to 206 and 2 up- or down-regulated genes. These genes were functionally annotated and grouped into the 5 categories described above, showing that the com-

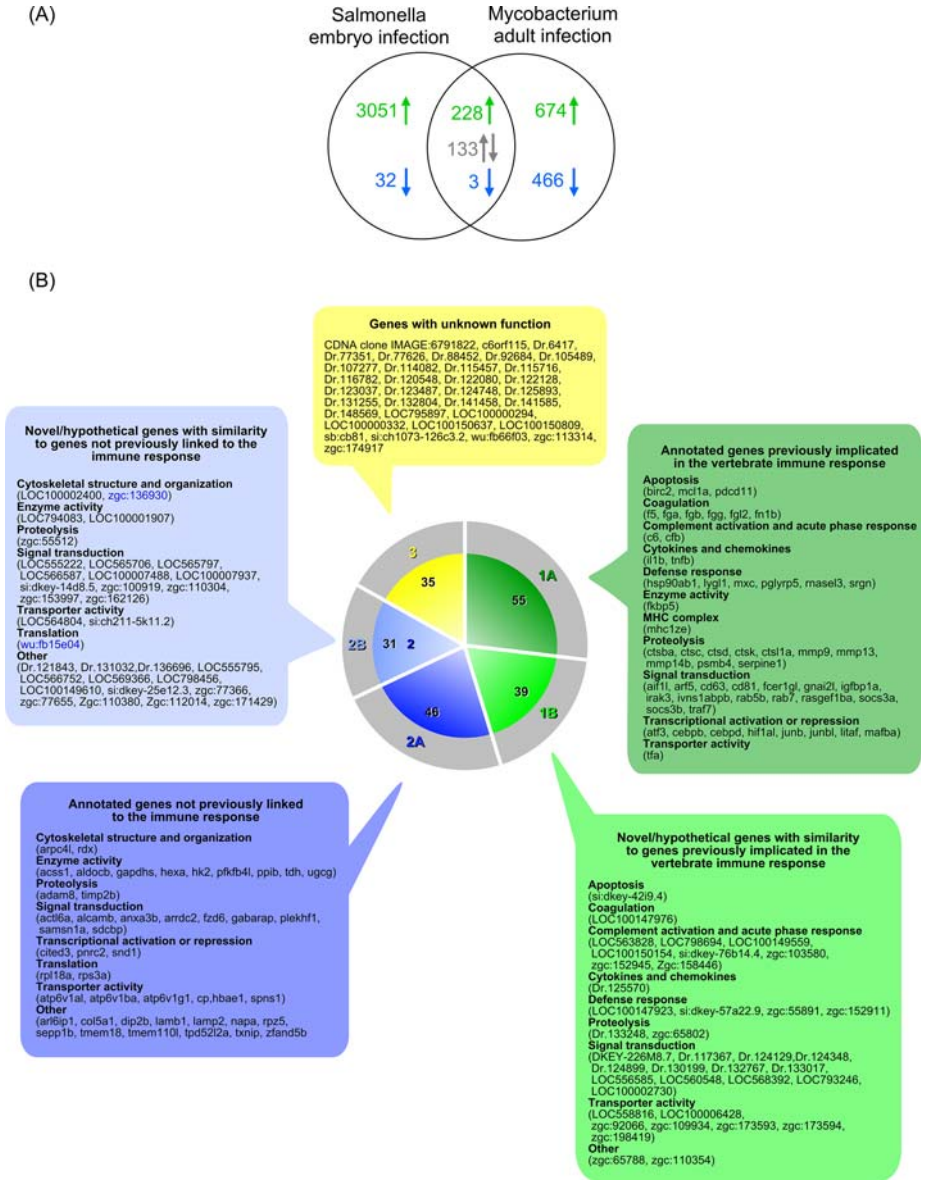

Figure 5. Comparison of infection responses in adult and embryonic zebrafish. The Tag-Seq data of *Salmonella* infection of zebrafish embryos reported here were compared with previously published DGE data of the end stage of *Mycobacterium* infection in adult zebrafish (Hegedus et al., 2009). (A) Venn diagram showing the overlap between DGE data of the different infection studies based on sCTDI calculation. In the overlapping sector of the Venn-diagram the numbers of up-regulated transcripts are indicated in green, numbers of down-regulated transcripts in blue, and the numbers of transcripts changed in opposite direction in the two different infection studies is

indicated in grey. (B) Annotation of the overlap group of up-regulated genes in the different infection studies. Genes were grouped into 5 categories as in Figure 4. The number of genes in each category is indicated in the pie diagram and the corresponding genes are ordered by functional groups. Genes up-regulated by *Salmonella* and *Mycobacterium* infection (206 genes) are indicated in black, and genes down-regulated by both infections (2 genes) are indicated in blue. Gene descriptions and accession numbers are given in Supplementary Table 2.

monly up-regulated genes included 94 annotated or novel/hypothetical genes with homology to human immune-related genes (categories 1A/B), 77 annotated or novel/hypothetical genes that have not been directly implicated in host defense in previous studies (categories 2A/B), and 35 genes for which we could make no functional prediction (category 3). Only two genes were identified as commonly down-regulated between the different infection studies, encoding a predicted intermediate filament protein and a ribosomal protein. Commonly up-regulated genes with previous links to host defense (categories 1A/B) were mainly associated with apoptosis, coagulation, complement activation and acute phase response, cytokine and chemokine activity, defense response, proteolysis, signal transduction, transcriptional activation and repression, and transporter activity. Commonly up-regulated genes in categories 2A/B were mainly associated with cytoskeletal structure and organization, enzyme activity, proteolysis, signal transduction, transcriptional activation and repression, translation and transporter activity.

Discussion

In this study we used the *S. typhimurium*-zebrafish infection model for a deep sequencing analysis of the embryonic innate host response to an inflammatory infection. Using both tag-based (Tag-Seq) and full transcript (RNA-Seq) sequencing approaches we extended and validated previous microarray data of this infection model (Stockhammer et al., 2009). The combined sequencing-based and microarray-based transcriptome data resulted in an annotated reference set of *Salmonella*-responsive genes in zebrafish embryos, including those homologous to human immune-related genes as well as many known or novel genes not previously linked to the immune response. Furthermore, comparison of the deep sequencing data of *Salmonella* infection in zebrafish embryos with previous deep sequencing data of *Mycobacterium* infection in adult zebrafish (Hegedus et al., 2009), defined a common set of innate host defense genes that are expressed both in the absence and presence of a fully developed adaptive immune system.

For deep sequencing analysis we chose *S. typhimurium* infected 1-day-old zebrafish embryos at 8 hpi, which time point was based on the strong induction of inflammatory genes that we detected in our previous microarray analysis (Stockhammer et al., 2009). Furthermore, we previously showed that induction of inflammatory genes such as *il1b* and *mmp9* relied on MyD88-dependent signaling

at this time point (Stockhammer et al., 2009). Consistent with the microarray results, more up-regulated than down-regulated transcripts were detected by Tag-Seq and RNA-Seq deep sequencing. However, for reasons currently unknown, the percentage of down-regulated transcripts was lower in Tag-Seq (3%) than in RNA-Seq (17%) and microarray analysis (26 %). Gene ontology analysis of the up-regulated gene sets of all three transcriptome analysis methods showed enrichment of the GO-term response to stimulus. Although less than 20% of the genes previously found to be up-regulated in microarray analysis could be confirmed by deep sequencing, a commonly responsive set of 165 transcripts could be defined encoding transcription factors, signal transduction proteins, cytokines and chemokines, complement factors, proteins involved in apoptosis and proteolysis, and proteins with anti-microbial activities. The overlap between Tag-Seq and RNA-Seq was slightly higher (241 transcripts) and confirmed the differential expression of these gene groups. The total overlap between the different transcriptome profiling methods might seem limited. As previously discussed (Hegedus et al., 2009; t Hoen et al., 2008), this can be attributed not only to technical differences, but also to differences in data processing and statistical evaluation, or to differences between the methods in discriminating between expression of different transcript isoforms. Importantly, the Tag-Seq and RNA-Seq data sets contained many differentially expressed transcripts that were not included in the microarray platform, and these data sets are therefore highly useful to extend and improve the microarray design.

In this study we only analyzed sequence reads mapping to UniGene, RefSeq or Ensembl transcript databases. However, about one third of the significant Tag-Seq reads did not map to transcript databases. Similarly, transcript mapping failed for one third of the RNA-Seq reads, even when one mismatch was allowed. While polymorphisms and reads extending over intron boundaries may at least partly account for mapping failures, these observations suggests that a large set of infection-responsive genes is still unknown, which can be of great interest for further studies.

We have previously shown that Tag-Seq data can be used to detect selective induction or repression of different transcript isoforms generated by alternative splicing, alternative polyadenylation or alternative transcription initiation (Hegedus et al., 2009). Such events may be detected when different tags for the same transcript show significant changes in opposite directions. In the present study we did not exploit this aspect of the deep sequencing technique, since our main objective was to define a robust marker set of inflammatory genes as a reference for infection studies in the zebrafish embryo model. For this reason, we focused specifically on those transcripts whose corresponding tags consistently changed in the same direction and developed the Cumulative Transcript Detection Index (CTDI). This index reflects the accumulated information from all tags that map to the same transcript, giving increased weight to tags with higher significance in statistical testing. The CTDI calculation proved useful for the comparison of Tag-Seq with microarray and RNA-Seq and increased the overlap between the data sets by approximately 30% in both cases as

compared to the use of the conventional Bayesian statistical evaluation method for the analysis of tag sequence data (Lash et al., 2000).

Previous studies (Mortazavi et al., 2008; t Hoen et al., 2008) have estimated that RNA-Seq analysis requires in the order of 10-fold more sequence reads for accurate quantification of expression differences between samples. Here, for the first time, a comparative Tag-Seq and RNA-Seq analysis of the host response to infection was performed. We found that 10 million Tag-Seq reads and 15 million RNA-Seq reads were both sufficient to detect around 1000 differentially expressed Ensembl transcripts during infection based on Bayesian statistical evaluation (Lash et al., 2000). Furthermore, Pearson correlation coefficients (>0.8) showed a linear relationship between the sequence read counts per Ensembl transcript in Tag-Seq and RNA-Seq, indicating the comparable performance of both methods in quantifying the transcriptome response to infection. Clearly RNA-Seq data have proved superior to Tag-Seq data for unraveling transcriptional landscapes (Wang et al., 2009). An advantage of Tag-Seq as compared to the RNA-Seq method used here, is that it allows discrimination between the expression of sense and antisense transcripts, which is of interest in view of the increasing evidence for the widespread occurrence and biological relevance of antisense transcription (Beiter et al., 2009; Carninci et al., 2005; Katayama et al., 2005). A substantial proportion of the differentially expressed sequence tags (10%) mapped to the antisense strands of known or predicted transcripts; however, this was much lower than previously observed in Tag-Seq analysis of *M. marinum* infection in adult zebrafish (40%; Hegedus et al., 2009).

Similar to *S. typhimurium* infection of zebrafish embryos, the end stage of *M. marinum* infection of adult zebrafish is also associated with a strong inflammatory response (Meijer et al., 2005; van der Sar et al., 2009). The similarity between these inflammatory responses at the level of gene expression was demonstrated here by an overlap of 206 up-regulated genes between the Tag-Seq data of *S. typhimurium* infection and the previously reported Tag-Seq data of *M. marinum* infection (Hegedus et al., 2009). This common set of infection-responsive genes included transcription factors and signaling components involved in the innate host defense, as well as genes not previously linked to the immune response of interest for further study in zebrafish models. The transcriptome data of both infection models provide a valuable reference for future studies of host-pathogen interactions in zebrafish.

Materials and methods

DGE (Tag-Seq) library construction and sequencing

The RNA samples for DGE analysis were identical to those used in Stockhammer et al., 2009. In brief, zebrafish embryos were infected with *Salmonella typhimurium* (strain SL1027) by microinjection of DsRED-labeled bacteria into the caudal vein close to the urogenital opening after the onset of blood circulation (27 hpf). An equal volume of PBS was injected in the control group. RNA samples were collected at 8

hours post infection (hpi) and samples from triplicate infection experiments were pooled. DGE libraries from the RNA pools (1 μg) of *Salmonella*-infected and control embryos were prepared using the DGE:Tag Profiling for NlaIII Sample Prep kit from Illumina as previously described (Hegedus et al., 2009). The libraries were sequenced in duplicate using 2 and 3 pmol of cDNA. Sequencing was performed using the Illumina Genome Analyzer II System (BaseClear B.V., Leiden, The Netherlands) according to the manufacturer's protocols. Image analysis, base calling, extraction of 17 bp tags and tag counting were performed using the Illumina pipeline. Tag counts from duplicate libraries were merged *in silico*.

RNA-Seq library construction and sequencing

Samples used for full RNA sequencing were the infected and uninfected control groups from a morpholino knock-down study to be reported elsewhere (Stockhammer et al., unpublished results). The procedure of *S. typhimurium* infection and the time-point of analysis (8 hpi) were identical as for the DGE analysis and previous microarray study (Stockhammer et al., 2009). Total RNA was isolated using the Qiagen miRNeasy kit according to the manufacturer's instructions (QIAGEN GmbH, Hilden). RNA-Seq libraries were made from 4 μg of each sample, using the Illumina mRNA-Seq Sample Preparation Kit according to the manufacturer's instructions (Illumina, Inc. San Diego). An amount of 4 pmol of each library was sequenced in one lane with a read length of 51 nt using the Illumina Genome Analyzer II System (BaseClear B.V., Leiden, the Netherlands).

DGE (Tag-Seq) data analysis

Mapping of tag sequences to transcript databases or to the zebrafish genome was performed as previously described (Hegedus et al., 2009). For transcript mapping we used the Ensembl *Danio rerio* Zv8.55 database, the RefSeq database (2009-09-14), and the *Danio rerio* UniGene build 105 and 117 databases. For comparison of Tag-Seq and RNA-Seq data the Ensembl transcript database derived from the latest version of the zebrafish genome was used. For comparison of Tag-Seq and microarray data we used the UniGene build 105 database, since this database was used in previous microarray analysis (Stockhammer et al., 2009). For genomic mapping the native and masked form of the zebrafish genome version Zv8 were downloaded from the FTP server of the Ensembl database. Statistical comparison of DGE/Tag-Seq data from *Salmonella*-infected and control embryos was performed using the Bayesian method described by Lash et al. (2000) with the software tool available from the SAGEmap resource (Lash et al., 2000). Briefly, the method performs a key-by-key comparison of two key-count distributions by generating a probability that the frequency of any key in the distribution differs by more than a given fold factor from the other distribution. For two Tag-Seq libraries, the algorithm performs a differential, tag-by-tag count comparison, with correction for the total size of the library. In our analysis we used a 2-fold factor difference of transcript expression level as the subject

of the Bayesian statistical evaluation. The algorithm returns a probability value (*P*) for each tag describing the chance that the detected count numbers represent a fold difference of the tag concentration between the investigated samples greater than or equal to 2. The change of a tag expression was accepted as significant if *P* was above 0.95. As an alternative means to evaluate differential expression in Tag-Seq datasets we developed the Cumulative Transcript Detection Index (CTDI), which accumulates data from all tags that map to the same transcript:

$$\sum_{i=0}^{n} D_i P_i^2 \left| \frac{\sum_{i=0}^{n} D_i P_i^2}{\sum_{i=0}^{n} \left| D_i P_i^2 \right|} \right|$$

where: *n* is the number of the detected tag entities in a transcript
P is the significance of tags (Lash et al., 2000)
D is the coefficient for the direction of change (1 for increase or -1 for de crease).

In the CTDI calculation P^2 is used for giving increased weight for tags with higher significance, while the formula gives lower weight to those transcripts where tags are present that show changes in the opposite direction. In other words, the CTDI measure reflects the extent of experimental confidence regarding the direction of the transcript expression change after infection. CTDI values were calculated separately for tags mapping to the sense strand (sCTDI) and the antisense strands (asCTDI) of transcripts in the database. All transcripts represented by minimally one tag with a significant expression change based on statistical evaluation according to Lash et al. (2000) have an absolute CTDI value larger than 0.95. For comparison with microarray data we considered all transcripts with an absolute sCTDI value larger than 0.95 and *P* larger than 0.8 for at least one of the tags in the CTDI calculation.

RNA-Seq data analysis

Sequence reads were mapped to Ensembl transcripts (Zv8.55) using the CLCbio Genomics Workbench version 3.6.5 (www.clcbio.com). RPKM values (read counts corrected for library size and transcript length) were calculated according to Mortazavi et al. (2008). Differential expression between control and infected samples was calculated based on total read counts per transcript using the method of Lash et al. (2000) as described above under Tag-Seq analysis.

Gene annotation

Gene Ontology (GO) analysis of significantly expressed *Salmonella*-responsive genes was performed using the GeneTools *e*GOn v2.0 web-based gene ontology analysis software (www.genetools.microarray.ntnu.no) at the level of the UniGene clusters (*D. rerio* UniGene build no. 105) (Beisvag et al., 2006). Overlapping genes

of Tag-Seq, microarray and RNA-Seq datasets were functionally annotated using the National Center for Biotechnology Information (NCBI) Gene Ontology Annotation (GOA), Entrez Gene, and HomoloGene databases and the AmiGO (http://amigo.geneontology.org/cgi-bin/amigo/go.cgi) gene ontology database. Human homologues of zebrafish genes were identified by Ensembl's BioMart data mining tool (http://www.ensembl.org/biomart/martview/) and using the NCBI HomoloGene database. The previous implication of genes in the vertebrate immune response was based on the GO data of the zebrafish genes and their human homologues supplemented by PubMed abstract searches and on overlap with the common host response gene list defined by Jenner and Young (Jenner and Young, 2005).

Supplementary data

Supplementary data associated with this chapter can be found at http://apo.szbk.u-szeged.hu/transfer/A_ORDAS/.

References

Beisvag V., Junge F. K., Bergum H., Jolsum L., Lydersen S., Gunther C. C., Ramampiaro H., Langaas M., Sandvik A. K. and Laegreid A. (2006) GeneTools--application for functional annotation and statistical hypothesis testing. *BMC Bioinformatics* **7**, 470.

Beiter T., Reich E., Williams R. W. and Simon P. (2009) Antisense transcription: a critical look in both directions. *Cell Mol Life Sci* **66**, 94-112.

Carninci P., Kasukawa T., Katayama S., Gough J., Frith M. C., Maeda N., Oyama R., Ravasi T., Lenhard B., Wells C., Kodzius R., Shimokawa K., Bajic V. B., Brenner S. E., Batalov S., Forrest A. R., Zavolan M., Davis M. J., Wilming L. G., Aidinis V., Allen J. E., Ambesi-Impiombato A., Apweiler R., Aturaliya R. N., Bailey T. L., Bansal M., Baxter L., Beisel K. W., Bersano T., Bono H., Chalk A. M., Chiu K. P., Choudhary V., Christoffels A., Clutterbuck D. R., Crowe M. L., Dalla E., Dalrymple B. P., de Bono B., Della Gatta G., di Bernardo D., Down T., Engstrom P., Fagiolini M., Faulkner G., Fletcher C. F., Fukushima T., Furuno M., Futaki S., Gariboldi M., Georgii-Hemming P., Gingeras T. R., Gojobori T., Green R. E., Gustincich S., Harbers M., Hayashi Y., Hensch T. K., Hirokawa N., Hill D., Huminiecki L., Iacono M., Ikeo K., Iwama A., Ishikawa T., Jakt M., Kanapin A., Katoh M., Kawasawa Y., Kelso J., Kitamura H., Kitano H., Kollias G., Krishnan S. P., Kruger A., Kummerfeld S. K., Kurochkin I. V., Lareau L. F., Lazarevic D., Lipovich L., Liu J., Liuni S., McWilliam S., Madan Babu M., Madera M., Marchionni L., Matsuda H., Matsuzawa S., Miki H., Mignone F., Miyake S., Morris K., Mottagui-Tabar S., Mulder N., Nakano N., Nakauchi H., Ng P., Nilsson R., Nishiguchi S., Nishikawa S., et al. (2005) The transcriptional landscape of the mammalian genome. *Science* **309**, 1559-63.

Davidson A. J. and Zon L. I. (2004) The 'definitive' (and 'primitive') guide to zebrafish hematopoiesis. *Oncogene* **23**, 7233-46.

Hegedus Z., Zakrzewska A., Agoston V. C., Ordas A., Racz P., Mink M., Spaink H. P. and Meijer A. H. (2009) Deep sequencing of the zebrafish transcriptome response to mycobacterium infection. *Mol Immunol* **46**, 2918-30.

Herbomel P., Thisse B. and Thisse C. (1999) Ontogeny and behaviour of early macrophages in the zebrafish embryo. *Development* **126**, 3735-45.

Jenner R. G. and Young R. A. (2005) Insights into host responses against pathogens from transcriptional profiling. *Nat Rev Microbiol* **3**, 281-94.

Katayama S., Tomaru Y., Kasukawa T., Waki K., Nakanishi M., Nakamura M., Nishida H., Yap C. C., Suzuki M., Kawai J., Suzuki H., Carninci P., Hayashizaki Y., Wells C., Frith M., Ravasi T., Pang K. C., Hallinan J., Mattick J., Hume D. A., Lipovich L., Batalov S., Engstrom P. G., Mizuno Y., Faghihi M. A., Sandelin A., Chalk A. M., Mottagui-Tabar S., Liang Z., Lenhard B. and Wahlestedt C. (2005) Antisense transcription in the mammalian transcriptome. *Science* **309**, 1564-6.

Lam S. H., Chua H. L., Gong Z., Lam T. J. and Sin Y. M. (2004) Development and maturation of the immune system in zebrafish, Danio rerio: a gene expression profiling, in situ hybridization and immunological study. *Dev Comp Immunol* **28**, 9-28.

Lash A. E., Tolstoshev C. M., Wagner L., Schuler G. D., Strausberg R. L., Riggins G. J. and Altschul S. F. (2000) SAGEmap: a public gene expression resource. *Genome Res* **10**, 1051-60.

Meeker N. D. and Trede N. S. (2008) Immunology and zebrafish: spawning new models of human disease. *Dev Comp Immunol* **32**, 745-57.

Meijer A. H., Gabby Krens S. F., Medina Rodriguez I. A., He S., Bitter W., Ewa Snaar-Jagalska B. and Spaink H. P. (2004) Expression analysis of the Toll-like receptor and TIR domain adaptor families of zebrafish. *Mol Immunol* **40**, 773-83.

Meijer A. H., Verbeek F. J., Salas-Vidal E., Corredor-Adamez M., Bussman J., van der Sar A. M., Otto G. W., Geisler R. and Spaink H. P. (2005) Transcriptome profiling of adult zebrafish at the late stage of chronic tuberculosis due to Mycobacterium marinum infection. *Mol Immunol* **42**, 1185-203.

Morrissy A. S., Morin R. D., Delaney A., Zeng T., McDonald H., Jones S., Zhao Y., Hirst M. and Marra M. A. (2009) Next-generation tag sequencing for cancer gene expression profiling. *Genome Res* **19**, 1825-35.

Mortazavi A., Williams B. A., McCue K., Schaeffer L. and Wold B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**, 621-8.

Salazar-Gonzalez R. M. and McSorley S. J. (2005) Salmonella flagellin, a microbial target of the innate and adaptive immune system. *Immunol Lett* **101**, 117-22.

Santos R. L., Zhang S., Tsolis R. M., Kingsley R. A., Adams L. G. and Baumler A. J. (2001) Animal models of Salmonella infections: enteritis versus typhoid fever. *Microbes Infect* **3**, 1335-44.

Stein C., Caccamo M., Laird G. and Leptin M. (2007) Conservation and divergence of gene families encoding components of innate immune response systems in zebrafish. *Genome Biol* **8**, R251.

Stockhammer O. W., Zakrzewska A., Hegedus Z., Spaink H. P. and Meijer A. H. (2009) Transcriptome profiling and functional analyses of the zebrafish embryonic innate immune response to Salmonella infection. *J Immunol* **182**, 5641-53.

t Hoen P. A., Ariyurek Y., Thygesen H. H., Vreugdenhil E., Vossen R. H., de Menezes R. X., Boer J. M., van Ommen G. J. and den Dunnen J. T. (2008) Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. *Nucleic Acids Res* **36**, e141.

Traver D., Herbomel P., Patton E. E., Murphey R. D., Yoder J. A., Litman G. W., Catic A., Amemiya C. T., Zon L. I. and Trede N. S. (2003) The zebrafish as a model organism to study development of the immune system. *Adv Immunol* **81**, 253-330.

Trede N. S., Langenau D. M., Traver D., Look A. T. and Zon L. I. (2004) The use of zebrafish to understand immunity. *Immunity* **20**, 367-79.

van der Sar A. M., Musters R. J., van Eeden F. J., Appelmelk B. J., Vandenbroucke-Grauls C. M. and Bitter W. (2003) Zebrafish embryos as a model host for the real time analysis of Salmonella typhimurium infections. *Cell Microbiol* **5**, 601-11.

van der Sar A. M., Spaink H. P., Zakrzewska A., Bitter W. and Meijer A. H. (2009) Specificity of the zebrafish host transcriptome response to acute and chronic mycobacterial infection and the role of innate and adaptive immune components. *Mol Immunol* **46**, 2317-32.

van der Sar A. M., Stockhammer O. W., van der Laan C., Spaink H. P., Bitter W. and Meijer A. H. (2006) MyD88 innate immune function in a zebrafish embryo infection model. *Infect Immun* **74**, 2436-41.

Wang Z., Gerstein M. and Snyder M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10**, 57-63.

Willett C. E., Cortes A., Zuasti A. and Zapata A. G. (1999) Early hematopoiesis and developing lymphoid organs in the zebrafish. *Dev Dyn* **214**, 323-36.